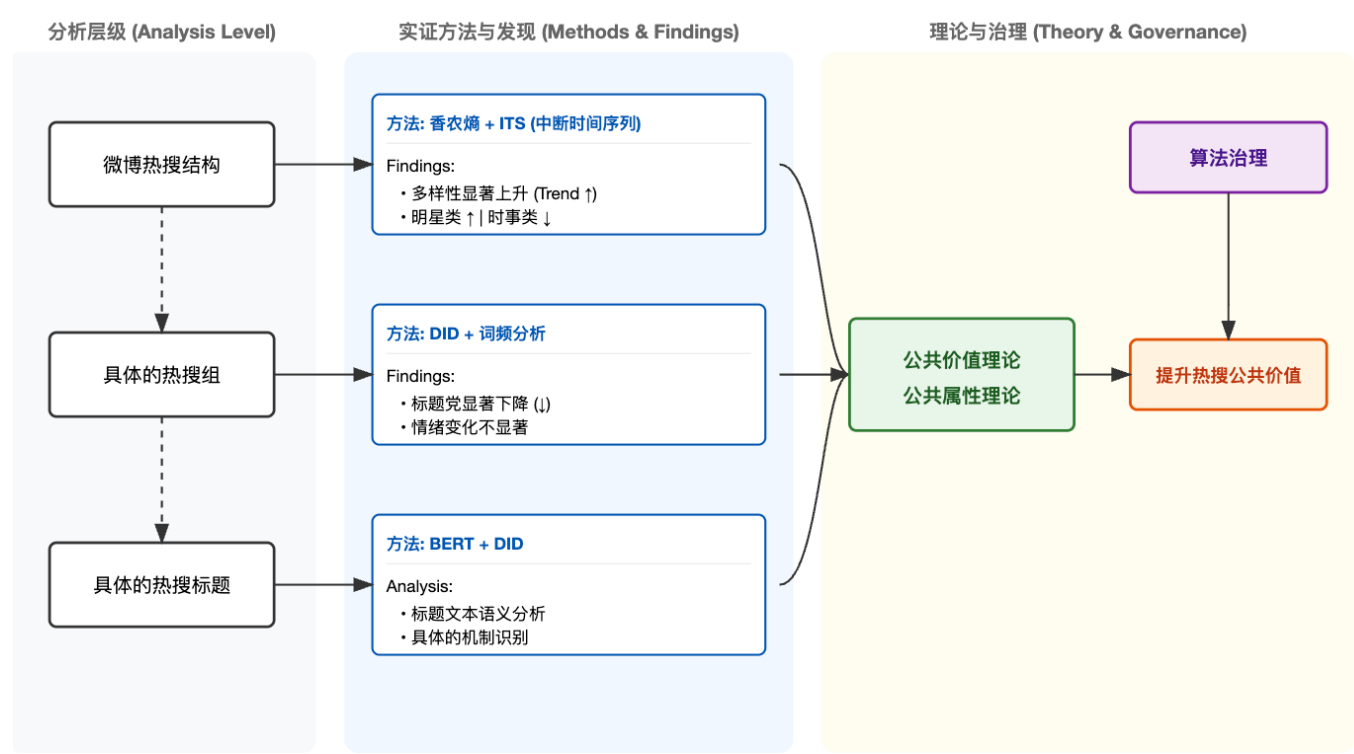
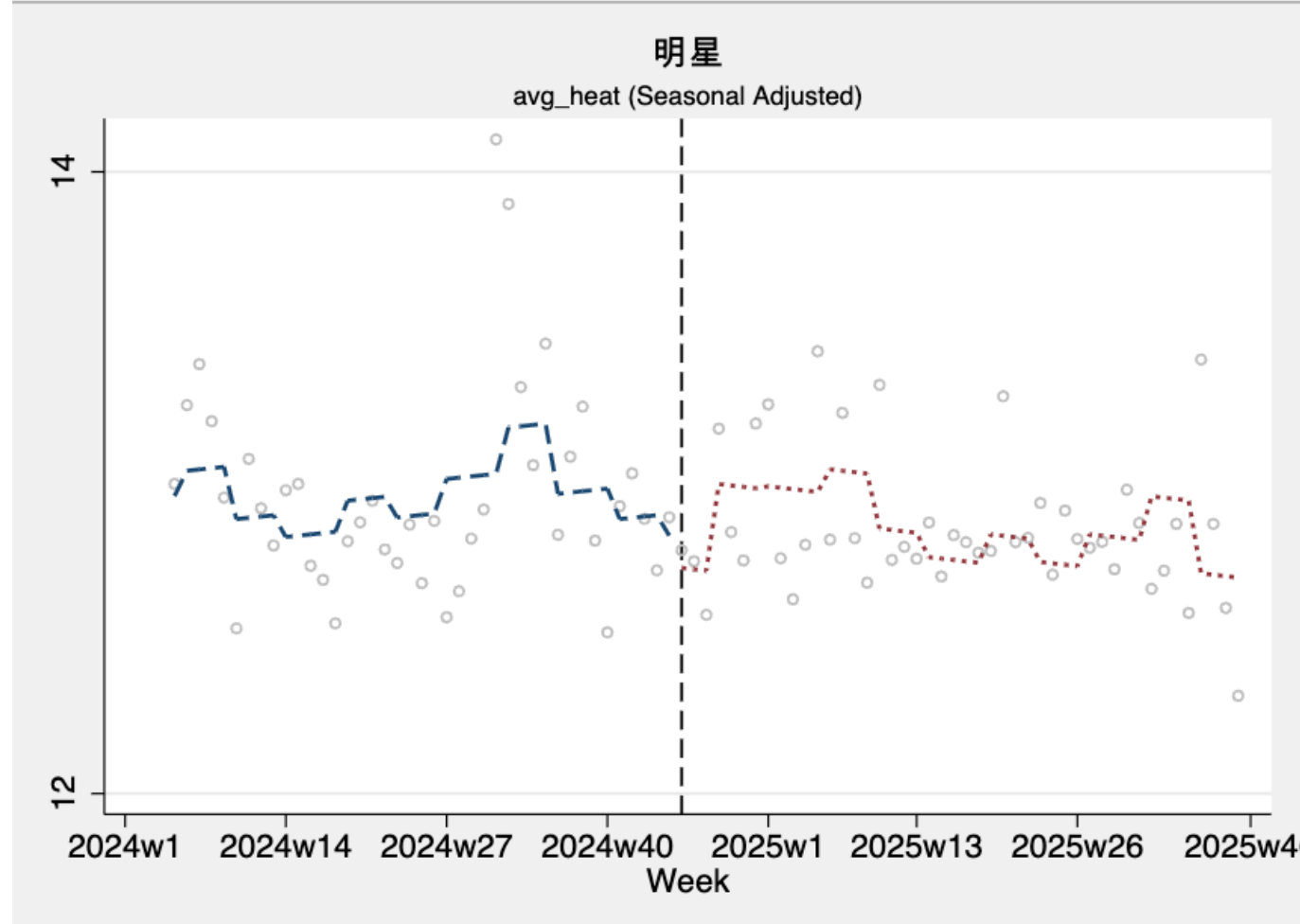


微博热搜治理机制与公共价值研究框架



![ITS结果]

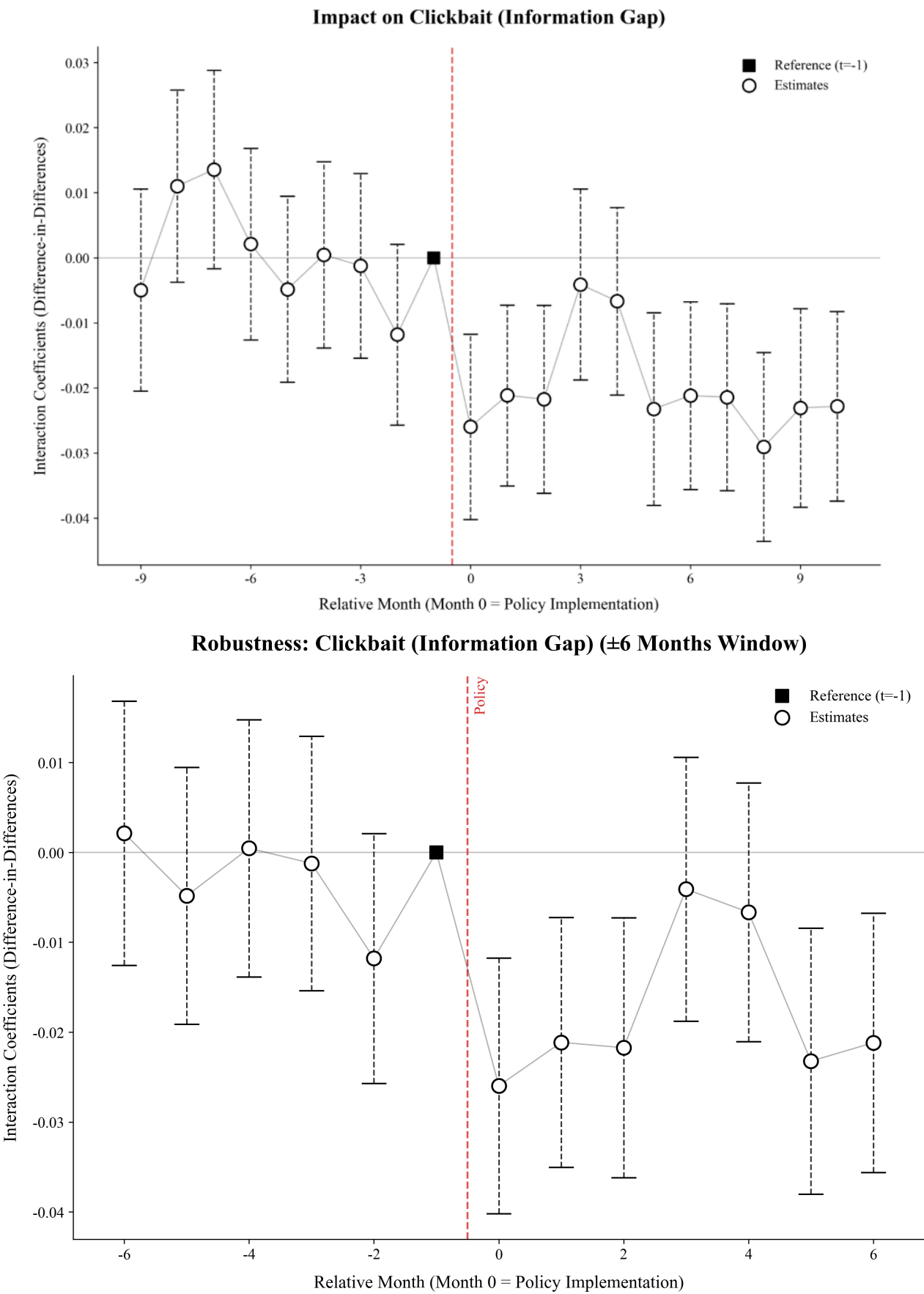


按 outcome 观察

- **total_heat / heat_per_mau / heat_per_dau**: 总体水平偏正 (均值 ≈ 0.159)，但趋势显著下行 (均值 ≈ -0.025 , 7/10 分类显著)。明星类水平提升最强 ($\approx +0.623$, $p \approx 0.0046$)，时事类水平下降明显 (≈ -0.439 , $p \approx 0.034$)。体育、财经斜率下行最陡。
- **avg_heat**: 整体水平为负 (均值 -0.195)，时事显著下跌 (-0.484 , $p \approx 0.005$)。趋势也下行 (-0.0104 , 4/10 显著)。
- **topic_count**: 水平明显上升 (均值 0.351 , 6/10 显著)；明星、财经拉升最大 ($+0.730$ 、 $+0.716$)。趋势方面仅“其他”小幅上行 ($+0.0138$, $p \approx 0.012$)，时事显著下降 (-0.0386 , $p \approx 4e-11$)。
- **normalized_heat**: 水平接近零 (均值 -0.0014)，但正负均显著 (明星 $+0.036$, 时事 -0.038 , p 均 $< 1e-4$)。趋势对“其他”显著上行、“社会”显著下行，幅度虽小但统计显著。

按分类观察

- **明星/明星-内地/游戏/财经**: 水平冲击普遍为正 (均值分别 0.421 、 0.215 、 0.345 、 0.379)，其中明星、游戏有 4-5 个显著水平提升；趋势多为小幅下行但幅度有限。
- **体育**: 水平变化不显著 (均值 0.133 , 无显著项)，但趋势显著下行最明显 (6/10 显著, 斜率均值 -0.035)，需关注持续衰减。
- **时事**: 水平与趋势均显著为负 (水平均值 -0.299 , 斜率均值 -0.028 , 显著项均 5 个)，是受冲击最明显的板块。
- **综艺/电视剧/社会**: 水平略为负或接近零，显著性有限；综艺与电视剧趋势接近持平，社会趋势显著下行 (5/10)。



BERT:

分类1:国家、社会、对立内容等, DID不显著

分类2:低、中、高社会价值: 使用 hfl/chinese-macbert-large 模型, 基于上述 sample 数据进行了微调训练。 *
公共价值分类 (Public Value): 验证集准确率约 80.36% * 公共属性分类 (Public Attribute): 验证集准确率约 82.14% * 模型已对 social_full.csv 完成全量预测。

考虑理论

- 公共价值理论
- 算法问责理论&信息完整性理论
- 提出了对于算法推荐审计的方法论
- 揭示了算法治理的直接效应和非预期效应